Proceedings of the ACM CHI 2000 Conference on Human Factors in Computing Systems (The Hague, The Netherlands, April 1-6, 2000), pp. 345-352. ACM Press, New York. Preprint version

Measuring Usability: Are Effectiveness, Efficiency, and Satisfaction Really Correlated?

Erik Frøkjær
Dept. of Computing
University of Copenhagen
Copenhagen Ø, Denmark
+45 3532 1456
erikf@diku.dk

Morten Hertzum
Centre for Human-Machine Interaction
Risø National Laboratory
Roskilde, Denmark
+45 4677 5145
morten.hertzum@risoe.dk

Kasper Hornbæk
Dept. of Computing
University of Copenhagen
Copenhagen Ø, Denmark
+45 3532 1452
kash@diku.dk

ABSTRACT
Usability comprises the aspects effectiveness, efficiency, and satisfaction. The correlations between these aspects are not well understood for complex tasks. We present data from an experiment where 87 subjects solved 20 information retrieval tasks concerning programming problems. The correlation between efficiency, as indicated by task completion time, and effectiveness, as indicated by quality of solution, was negligible. Generally, the correlations among the usability aspects depend in a complex way on the application domain, the user's experience, and the use context. Going through three years of CHI Proceedings, we find that 11 out of 19 experimental studies involving complex tasks account for only one or two aspects of usability. When these studies make claims concerning overall usability, they rely on risky assumptions about correlations between usability aspects. Unless domain specific studies suggest otherwise, effectiveness, efficiency, and satisfaction should be considered independent aspect of usability and all be included in usability testing.

Keywords
Usability measures, effectiveness, efficiency, satisfaction, information retrieval, usability testing, user studies

INTRODUCTION
Although the importance of usability is gaining widespread recognition, considerable confusion exists over the actual meaning of the term. Sometimes usability is defined quite narrowly and distinguished from, for example, utility [11], on other occasions usability is defined as a broad concept synonymous to quality in use [2]. We adopt ISO's broad definition of usability [7] as consisting of three distinct aspects:

- *Effectiveness*, which is the accuracy and completeness with which users achieve certain goals. Indicators of effectiveness include quality of solution and error rates. In this study, we use quality of solution as the primary indicator of effectiveness, i.e. a measure of the outcome of the user's interaction with the system.
- *Efficiency*, which is the relation between (1) the accuracy and completeness with which users achieve certain goals and (2) the resources expended in achieving them. Indicators of efficiency include task completion time and learning time. In this study, we use task completion time as the primary indicator of efficiency.
- *Satisfaction*, which is the users' comfort with and positive attitudes towards the use of the system. Users' satisfaction can be measured by attitude rating scales such as SUMI [8]. In this study, we use preference as the primary indicator of satisfaction.

While it is tempting to assume simple, general relations between effectiveness, efficiency, and satisfaction, any relations between them seem to depend on a range of issues such as application domain, use context, user experience, and task complexity. For routine tasks good performance depends on the efficient, well-trained execution of a sequence of actions which is known to yield stable, high-quality results [3]. For such tasks high-quality results are routinely achieved, and task completion time may therefore be used as an indicator of overall usability. For non-routine, i.e. complex tasks, there is no preconceived route to high-quality results, and good performance is primarily dependent on conceiving a viable way of solving the task [9, 14]. The efficient execution of the sequence of actions is of secondary importance. Consequently, efficient execution of the actions may or may not lead to high-quality results, and diligence is not even guaranteed to lead to task completion. This suggests that, at least for complex tasks, efficiency measures are useless as indicators of usability unless effectiveness is controlled.

Nielsen & Levy [12] analyzed the relation between efficiency and user preference in 113 cases extracted from 57 HCI studies. Their general finding was that preference predicts efficiency quite well. However, in 25% of the cases the users did not prefer the system they were more efficient in using. The ambition of finding a simple, general relationship between efficiency and satisfaction is therefore questionable [see also 1]. Studies of, for example, specific application domains may yield more precise and informative models. With respect to the relationship between satisfaction and effectiveness, Nielsen & Levy [12] note that their very comprehensive literature survey did not encounter a single study that compared indicators of these two aspects of usability.

In this paper we investigate the connection between efficiency, indicated by task completion time, and effectiveness, indicated by quality of solution. This is done by reanalyzing data from the TeSS-experiment [6] where 87 subjects solved a number of information retrieval tasks, using four different modes of the TeSS system and programming manuals in hard copy. In analyzing the data we look for correlations between efficiency and effectiveness across retrieval modes, tasks, and individual subjects.

The purpose of this paper is to emphasize the importance of accounting for all three aspects of usability in studies that assess system usability, for example to compare the usability of different designs. Effectiveness is often difficult to measure in a robust way. This may be the reason why several studies involving complex tasks refrain from accounting for effectiveness and settle for measures of the efficiency of the interaction process [for example, 5, 13]. These studies rest on the assumption that an efficient interaction process indicates that the user also performed well in terms of crucial effectiveness indicators such as solution quality. The TeSS-experiment illustrates that this assumption is not warranted—unless it can be supported by an argument that effectiveness is controlled.

The first two sections present the method and results from the TeSS-experiment, establishing the argument that efficiency and effectiveness are weakly—if at all—correlated. Next, we discuss the general relationship between the three aspects of usability, exemplifying the impact of our findings by studies from the CHI Proceedings of the years 1997-99. We then discuss the implications of our findings with regard to the selection of usability measures. In the final section, we outline our main conclusions concerning the weak and context-dependent relation between the usability aspects.

THE TESS-EXPERIMENT

The purpose of the TeSS-experiment was to compare the usage effectiveness of browsing and different forms of querying in information retrieval tasks concerning programming problems. Further, the experiment aimed at establishing a detailed description of the subjects' interaction with the TeSS system.

Experimental Conditions

To solve the tasks the subjects needed information concerning the development of graphical user interfaces in the X Window System. Access to the necessary documentation (approximately 3 Mb of text) was provided through an experimental text retrieval system called TeSS and by means of manuals in hard copy. TeSS can be operated in four different modes, each providing the user with a different set of retrieval facilities. Thus, the experiment involves five retrieval modes:

- BROWSE. In TeSS, browsing can be done by expanding and collapsing entries in the table of contents and by searching the table of contents for specific strings. The text itself is presented in separate windows.
- LOGICAL. A mode of TeSS offering conventional Boolean retrieval where queries are logical expressions built of query terms, ANDs, ORs, NOTs, parentheses, and wildcards.
- VENN. In this mode of TeSS queries are expressed by means of a Venn diagram which replaces Boolean operators with a, supposedly, more immediately understandable graphical image of intersecting sets.
- ALL. The whole of TeSS offering the combination of BROWSE, LOGICAL, and VENN.
- PAPER. In this mode searching is done in hard copies of the programming manuals, i.e. independently of TeSS.

Subjects

The subjects were 87 students in their third year of a bachelor degree in computer science. While the project was a mandatory part of the students' education, participation in the experiment by allowing the data collection to take place was voluntary and anonymous. The subjects were first-time users of TeSS and had no prior knowledge of the programming tools on which the tasks were based.

Tasks

In the TeSS-experiment each subject solved 20 information retrieval tasks. As preparation, the subject completed two practice tasks. The 20 tasks concerned whether and how certain interface properties could be achieved in a graphical user interface. To answer the tasks the subjects had to identify the relevant user interface objects, e.g. widgets, methods, and resources, and outline an implementation. As the subjects were unfamiliar with the X Window System, the tasks involved a substantial element of learning in addition to the need for retrieving specific pieces of information. Some tasks were formulated in the context of the X Window System in general; others took the user

interface of TeSS as their point of departure. Two examples of tasks used in the TeSS-experiment are:

Task 5. Radio buttons are used in situations where exactly one option must be chosen from a group of options. Which widget class is used to implement radio buttons?

Task 11. The caption on the button "done" should be changed to "quit". How is that done?

Procedure

The experiment was explained to the subjects at a lecture, after which the subjects had ten days to complete the tasks. The subjects received a manual for TeSS and a two-page walk-up-and-use introduction. The system itself was available on terminals to which students have access 24 hours a day. The manual searching was done in the library where one of the authors was present three hours a day to hand out tasks and receive solutions. Upon entering the library, the subjects received hard copies of the three manuals, a sheet with the proper task, and a log sheet with fields for starting time, finishing time, and solution.

The experiment employed a within-groups design where all subjects solved the tasks in the same sequence and each subject was required to use all retrieval modes. To avoid order effects, the subjects were exposed to the retrieval modes in a systematically varied order. The 20 information retrieval tasks were clustered into five blocks. The first block was solved with one of the five retrieval modes, the second block with one of the remaining four retrieval modes. Thus the permutations of the modes on the two first blocks divided the subjects into 20 groups. The number of subjects did not allow all 5! sequences of the five modes to be included, and the 20 groups were not divided further. Rather, the order of the three remaining modes was kept the same within each group.

Data Collection and Analysis

The data collected in the experiment include a detailed log of the subjects' interaction with TeSS. The interaction log gives a time-stamped account of the commands executed by the subjects. It also includes task demarcation and solutions reached, both obtained from a separate module governing the subjects' access to TeSS. This Task Handling Module makes it possible to let the subjects work unsupervised while at the same time enforcing a strict experimental procedure. The Task Handling Module presents the tasks to the subject one at a time, gives access to the retrieval mode to be used by that subject when solving that particular task, and records his or her solution. For the PAPER retrieval mode, the subjects recorded their starting time, finishing time, and task solution on the log sheets.

The 87 subjects received 20 information retrieval tasks each, giving a potential total of 1740 answers. However, 113 answers were not submitted; 19 were excluded because they included a more than one hour long period with no logged user activity; 17 were excluded due to technical problems with TeSS; 14 were excluded because it was impossible to judge the quality of the answer; and 2 were excluded because they were solved poorly in less than two minutes, i.e., without any attempt to reach a solution. Finally, 4 subjects were excluded because they clearly did not take the experiment seriously. Thus, 11% of the answers were not submitted or excluded. The analysis is based on the remaining 1555 answers, the results of 648 hours of work performed by 83 subjects.

Grade	Mnemonic	Description
1	Very low	Failure, a completely wrong answer
2	Low	Inadequate or partially wrong answer
3	Medium	Reasonable but incomplete answer
4	High	Good and adequate answer
5	Very high	Brilliant answer

Table 1—The five-point scale used to grade the tasks

In this paper we focus on two aspects of the usability of TeSS:

- Efficiency measured as task completion time, which is extracted from the interaction log or the log sheets.
- Effectiveness measured as the quality of the solution, which was assessed by one of the authors and expressed by a grade on a five-point scale, see Table 1. As an example, a medium and a high quality solution to task 5 (see above) must identify toggle widgets as the relevant widget class. A brilliant answer also explains the use of radio groups to cluster the toggle widgets.

The following analysis is restricted to the 20 information retrieval tasks—the bulk of our data. Data concerning user satisfaction, measured as subjects' preference for one or the other retrieval mode, were collected for three implementation tasks, which followed the information retrieval tasks. The preference data show that the subjects did not prefer the retrieval mode with which they performed best. Rather, they overwhelmingly preferred ALL, the retrieval mode where they did not exclude themselves from any of the search facilities available in BROWSE, BOOLEAN, or VENN [6]. This suggests that user satisfaction is not simply correlated with performance measures such as task completion time and grade. Thus, the TeSS-experiment was another exception to the general finding of Nielsen & Levy [12] that users prefer the objectively best system.

RESULTS OF THE TESS-EXPERIMENT

Table 2 shows the relation between task completion time and grade for the 1555 tasks solved in the TeSS-experiment. A contingency analysis of this table suggests that task completion time and grade are not independent (χ^2[16, N=1555]=47.81, p<0.001).

Task completion time Grade (no. of observations)	$<P_{20}$	P_{20}-P_{40}	P_{40}-P_{60}	P_{60}-P_{80}	$>P_{80}$	Mean time for grade (SD)
5 (N=147)	17	35	33	31	31	24.27 (20.62)
4 (N=566)	170	121	92	96	87	21.71 (38.80)
3 (N=216)	37	48	55	38	38	24.70 (26.18)
2 (N=192)	29	35	46	36	46	26.72 (32.60)
1 (N=434)	58	72	85	110	109	28.94 (27.35)
Median grade (P_{25}-P_{75})	4 (2-4)	4 (2-4)	3 (1-4)	3 (1-4)	3 (1-4)	

Table 2—Distribution of task completion time and grade for all tasks in the TeSS-experiment (N=1555). The column to the left shows the five grades given to the tasks, cf. Table 1. The next columns show the number of tasks in each of five intervals based on the 20, 40, 60, and 80 percentiles of task completion time. The rightmost column shows the mean time in minutes for a

Task completion time for subjects receiving a certain grade varies much, as can be seen from the large standard deviations in Table 2. An analysis of variance shows significant variation in task completion times between different grades ($F[4,1550]=3.31$, $p<0.01$). However, we did not find any pairwise differences between grades using Tukey's post hoc test at a five-percent significance level.

The tasks in any of the five intervals of task completion times shown in Table 2 received markedly different grades. Between time intervals there is significant variation in grades (analysis of variance with time interval as the independent and grade as the dependent variables, $F[4,1550]=9.10$, $p<0.001$). Pairwise comparisons of the five time intervals using Tukey's post hoc test show that the 20% fastest solved task receive significantly higher grades than the 60% slowest solved tasks. Similarly, solutions to tasks in the P_{20}-P_{40} time interval receive significantly higher grades than solutions in the time intervals P_{60}-P_{80} and $>P_{80}$.

Spearman's rank order correlation analysis shows that task completion time and grade are significantly correlated in tasks solved in the TeSS-experiment ($r_s=-0.156$, two-tailed p-level <0.001). Using more time for completing a task is thus correlated with receiving a lower grade. However, the correlation between time and grade is weak; only two percent of the variation in grade can be predicted from task completion time ($r_s^2=0.024$). According to [4] a correlation of this magnitude is negligible.

Retrieval mode (no. of observations)	Mean time (SD)	Median grade (P_{25}-P_{75})	r_s	p	r_s^2%
Browse (N=310)	22.88 (20.89)	3 (1-4)	-0.150	0.008	2.2
Logical (N=307)	30.15 (34.70)	3 (1-4)	-0.089	0.119	-
Venn (N=305)	25.79 (25.45)	3 (1-4)	-0.107	0.062	-
All (N=314)	30.80 (51.84)	3 (1-4)	-0.128	0.030	1.6
Paper (N=319)	15.66 (11.27)	4 (2-4)	-0.265	0.001	7.0

Table 3—Correlation between time and grade in different retrieval modes. The first column shows the retrieval modes, and the second and third columns the mean time in minutes and the median grade for each mode. Columns four to six show the Spearman correlation coefficient between time and grade r_s, the

To control for interplay between the design of the experiment and the weak correlation found, we performed a partial correlation analysis of the TeSS data. In the partial correlation analysis, the influence from different tasks and retrieval modes is removed from the correlation coefficient between time and grade [4]. This analysis also reveals a weak but statistically significant correlation between task completion time and grade (Spearman's partial correlation coefficient $r_s[\text{time,grade}|\text{configuration,task}]=-0.170$, $p<0.001$).

These analyses show that at the general level efficiency and effectiveness are only weakly correlated. In spite of this, time and grade could be correlated at a more detailed level of analysis, hereby undermining the conclusion at the general level. In the following sections we therefore analyze whether time and grade are correlated for specific retrieval modes, tasks, or subjects.

Correlation between Time and Grade for Different Retrieval Modes

The retrieval modes LOGICAL and VENN—the only retrieval modes requiring the subjects to formulate queries—do not show a significant correlation between time and grade (see Table 3). The retrieval modes BROWSE, ALL, and PAPER all show a statistically significant but weak correlation between task completion time and grade (r_s^2% between 1.6 and 7.0). The tasks solved in the retrieval mode PAPER have a numerically larger correlation between time and grade than the other retrieval modes. However, the correlation for PAPER is still weak and not significantly different from the correlations for

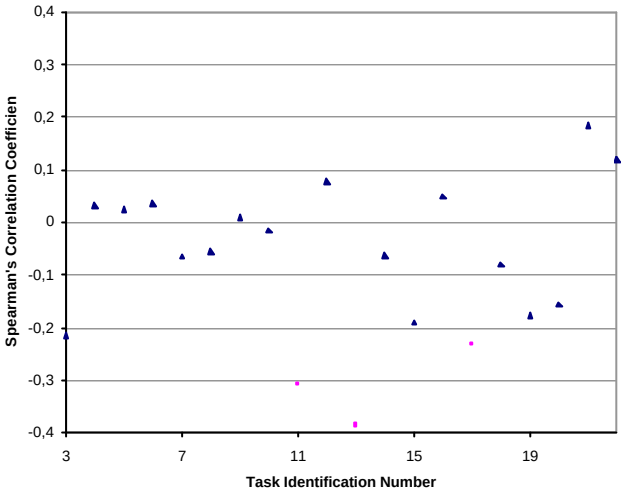

Figure 1—Correlation between time and grade for different tasks. The figure shows Spearman's correlation coefficient (r_s) for each of the 20 information retrieval tasks. Each task has been solved by between 69 and 81 subjects. Time and grade are significantly correlated for tasks 11, 13, and 17. These tasks appear as squares in the figure. The task identification numbers begin at 3, because tasks 1 and 2 are tasks used for training [6].

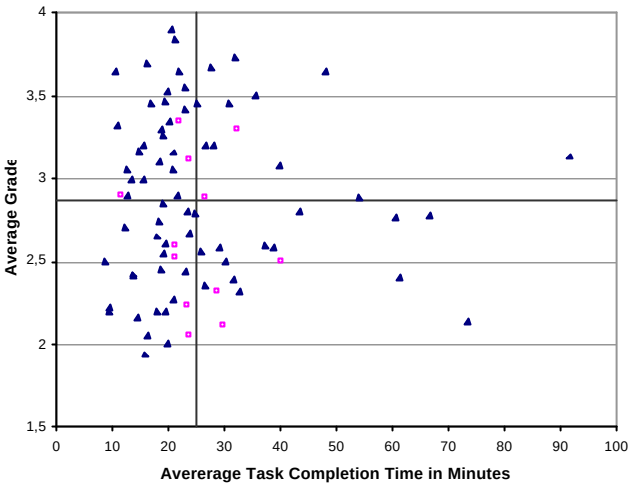

Figure 2—Average time and grade for each of the 83 subjects included in the data analysis. The horizontal line indicates the overall mean grade (2.87), the vertical line the overall mean time (25 min.). Subjects with a significant correlation between time and grade appear as squares, other subjects appear as triangles.

BROWSE and ALL (Fisher's r-to-z transformation, ALL vs. PAPER: z=-1.783, p>0.075, BROWSE vs. PAPER, z=-1.504, p>0.133).

Correlation between Time and Grade for Different Tasks

The correlation between task completion time and grade varies somewhat across the tasks (see Figure 1). For 85% of the tasks there is no correlation between time and grade. However, three tasks show a significant correlation between time and grade: task 11 (r_s=-0.308, p<0.007), task 13 (r_s=-0.387, p<0.001), and task 17 (r_s=-0.232, p<0.040). For these tasks between 5% and 15% of the variation in grade can be predicted from time, where more time spent is correlated with lower grade.

Task 11 and task 13 have a higher average grade than the other tasks (task 11: mean grade 3.42, t[1393]=-3.734, p<0.001; task 13: mean grade 3.72, t[1398]=-5.739, p<0.001). Task 13 is also solved faster than the other tasks (mean completion time 13.43 minutes, t[1398]=3.316, p<0.001). The description of these tasks given to the subjects specifies in detail some of the central interface objects of the tasks (see for example the wording of task 11 showed earlier). For task 17 it is only the relation between time and grade that is significant, individually neither time nor grade differs significantly from the other tasks.

Correlation between Time and Grade for Different Subjects

Looking at the average performance of subjects, the tasks solved by 12 of the subjects show a significant correlation between time and grade (see Figure 2). These correlation coefficients are all negative, suggesting that more time spent is correlated with lower grade (r_s between –0.758 and –0.453). For 86% of the subjects, time does not predict grade at all.

It is difficult to find a common denominator for the subjects where time and grade are correlated. The average time and grade of those subjects vary above and below the mean time and grade for subjects (see Figure 2). However, there is a significant difference between the grade for subjects with a significant correlation between time and grade and those without (Wilcoxon test, z=2.393, p<0.017). Subjects who obtain a correlation between time and grade did not use a specific retrieval mode for certain tasks (Chi-square test of which retrieval mode was first used, $?^2$[4, N=12]=3.833, p>0.05).

Summary of Correlations between Usability Measures

Our analysis of the TeSS-experiment shows that efficiency (measured as task completion time) and effectiveness (measured as grade) are either not correlated or correlated so weakly that the correlation is negligible for all practical purposes. For the individual retrieval modes, a weak correlation is found for three of the modes, while two of the modes do not show any significant correlation between task completion time and grade. Task completion time and grade are not correlated for 85% of the tasks. Finally, only 14% of the subjects display a significant correlation between time and grade—for the large majority no correlation is found. These results and the previous results [6] concerning satisfaction and effectiveness (cf. the section Data Collection and Analysis, last paragraph) show

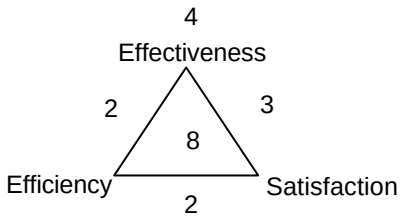

Figure 3—The usability aspects measured in the 19 studies of complex tasks from CHI '97 to CHI '99. Eight of these CHI-studies include measures of all three usability aspects, seven CHI-studies measure two aspects, and four CHI-studies only one aspect.

that assumptions about correlations between effectiveness, efficiency, and satisfaction do not seem to hold in the context of the TeSS-experiment.

CORRELATIONS BETWEEN ASPECTS OF USABILITY

We now extend the discussion of correlations between aspects of usability by including studies of computer support for complex tasks published in the CHI Proceedings for the years 1997-99. A total of 19 studies investigate aspects of usability in sufficient detail to enable an analysis of their choice of usability measures, see Figure 3. Eight (42%) of the 19 studies cover all three usability aspects. The other 11 studies, implicitly or explicitly, rely on assumptions of correlations between the different usability aspects, or seem confident that their choice of only one or two aspects of usability is sufficient to capture overall usability.

The only CHI-study with an analysis of correlations between the three aspects of usability

Of the eight studies including measures of all three usability aspects, only the study by Walker et al. [17] has analyzed the correlations between the aspects. Let us summarize their study, so the reader can see that the correlation analysis pays off.

Walker et al. compare two different designs of a spoken language interface to email: (a) a mixed-initiative dialogue, where the users can flexibly control the dialogue, and (b) a system-initiative dialogue, where the system controls the dialogue. The study measures effectiveness by qualitative measures such as automatic speech recognition rejects, efficiency by number of dialogue turns and task completion time, and user satisfaction by a multiple-choice survey. The results show that even though the mixed-initiative dialogue is more efficient, as measured by task completion time and number of turns, users prefer the system-initiative dialogue.

A correlation analysis with user satisfaction as the dependent variable uncovers how "…users' preferences are not determined by efficiency per se, as has been commonly assumed. One interpretation of our results is that users are more attuned to qualitative aspects of the interaction." [17, p. 587]. The number of automatic speech recognition rejects contributes the most to user satisfaction. Walker et al. suggest that the users' preference for the system-initiative dialogue arises from it being easier to learn and more predictable. This result was contrary to the authors' initial hypothesis and illustrates the importance of measuring efficiency, effectiveness, and satisfaction independently, as opposed to basing conclusions about one of them on measures of the others.

Two CHI-studies without any measure of effectiveness

Two CHI-studies concerning computer support for complex tasks, entitled "Time-compression: systems concerns, usage, and benefits" [13] and "Effects of awareness support on groupware usability" [5], do not include any measure of the quality of the outcome of the users' interaction with the system. Below we comment on these two studies, and show how their conclusions about overall usability are jeopardized by their incomplete choice of usability measures.

In the first study, Omoigui et al. [13] analyze how time-compression can be used to enable quick video browsing. An experimental time-compression system was used for comparing different granularities of the time-compression (discrete vs. continuous) and differences in the latency (long wait-time vs. no wait-time) experienced by users after adjusting the degree of time-compression. Omoigui et al. measure efficiency by savings in task time and the use of time-compression, and they measure satisfaction by, e.g., user feedback and preference indicated by usage of time-compression during video browse sessions. As already mentioned, no effectiveness measures were employed, although effectiveness could have been measured as the accuracy and completeness of the subjects' verbal summary of each video. In the concluding remarks, Omoigui et al. emphasize efficiency as the important aspect of time-compression systems: "Quite surprisingly though, there are no significant differences in the time-savings under the three conditions. Thus the implementers are free to choose the simplest solution..." [13, p. 142]. This conclusion neglects the satisfaction measures, which indicate that real differences might exist between the experimental conditions: "… several subjects commented in post-study debriefing that the long latency and discrete granularity conditions had affected their use of the time compression feature. The subjects felt that they made fewer adjustments and watched at a lower compression rate when long latency and discrete granularity were used." [13, p. 141]. An analysis of the correlations between the efficiency and satisfaction measures might have shed further light on the differences between conditions, as might solid measures of effectiveness.

In the second study, Gutwin and Greenberg [5] analyze whether enhanced support for workspace awareness improves collaboration. In an experiment, they compare users' performance on two real-time groupware systems where workspace miniatures were used to support workspace awareness. The basic miniature shows information only about the local user, the enhanced miniature about others in the workspace as well. Efficiency is measured by task completion time and communication efficiency; satisfaction is measured as preference for one or the other system. The correlations between the measures are not analyzed, and no measure of effectiveness is employed. The overall conclusion of the study is that workspace-awareness information reduces task completion time, and increases communicative efficiency and user satisfaction. The support for this conclusion is weak. For one out of the three task types, task completion time was not reduced. For two task types out of the three, the communicative efficiency was not increased. All 38 participants preferred the awareness-enhanced system, suggesting that the employed measures of usability are incomplete: "The overwhelming preference for the interface with the added awareness information also suggests that there were real differences in the experience of using the system, but that our measures were insensitive to these differences." [5, p. 517]. These differences might have been more explainable if the study had included measures of effectiveness, making possible an analysis of how users' preferences were affected by the quality of the outcome of their activities.

SELECTION OF USABILITY MEASURES

We believe that the weak correlation between effectiveness, efficiency, and satisfaction has three implications regarding the choice of measures in evaluations of system usability.

First, it is in general recommendable to measure efficiency, effectiveness as well as satisfaction. When researchers or developers use a narrower selection of usability measures for evaluating a system they either (a) make some implicit or explicit assumptions about relations between usability measures in the specific context, or (b) run the risk of ignoring important aspects of usability. In our analysis of the CHI-studies we have shown how interpretation of experimental data based on only one or two usability aspects leads to unreliable conclusions about overall usability. Given that the three usability aspects capture different constituents of usability—we have not seen arguments to the contrary for complex tasks—there is no substitute for including all three aspects in usability evaluations.

Second, at the moment no clear-cut advice can be given about which usability measures to use in a particular situation. On the contrary, identifying the usability measures that are critical in the particular situation should be recognized as a central part of any evaluation of system usability. This requires a firm understanding of how tasks, users, and technology interact in constituting the use situations within the particular application domain [10, 16]. The study by Su [15] is an illustrative example of the kind of work needed to distinguish and refine performance measures. Su investigated the correlation between 20 measures of information retrieval performance in an academic setting, and suggests a best single measure (the user's perception of the value of the search result as a whole) and best pairs of measures of information retrieval performance. Such work may lead to the development of reliable, domain-specific collections of critical performance measures. General descriptions of the relation between usability aspects [e.g. 12] will not aid the selection of usability measures, since there is no way of knowing in advance whether efficiency, effectiveness, and satisfaction are actually correlated in a particular situation.

Third, effectiveness measures oriented toward the outcome of the user's interaction with the system are gaining attention in usability evaluation [2], although two of the CHI-studies discussed earlier did not include such measures. The development of valid and reliable outcome measures is a prerequisite for assessing overall system usability and is necessary for working systematically with improving the usability of systems supporting users in solving complex tasks.

CONCLUSION

The relations between efficiency, effectiveness, and satisfaction—the three aspects of usability—are not well understood. We have analyzed data from a study of information retrieval and found only a weak correlation between measures of the three usability aspects. Other studies imply that for complex tasks in other domains, a similarly weak correlation between usability measures is to be expected. In general, we suggest that efficiency, effectiveness, and satisfaction should be considered independent aspects of usability, unless domain specific studies suggest otherwise.

Studies that employ measures of only a subset of the three usability aspects assume either that this subset is sufficient as an indicator of overall usability or that the selected measures are correlated with measures covering the other aspects of usability. As we have exemplified with an analysis of studies from previous CHI Proceedings, such assumptions are often unsupported. Hence, these studies jump to conclusions regarding overall usability while measuring, say, efficiency only. This is a problem for the HCI community, since more than half of the last three years of CHI-studies concerning complex tasks do not measure all aspects of usability.

Usability testing of computer systems for complex tasks should include measures of efficiency, effectiveness, and user satisfaction. In selecting these measures, the

application domain and context of use have to be taken into account so as to uncover the measures that are critical in the particular situation. Discovering solid measures of effectiveness seems especially critical.

ACKNOWLEDGEMENTS

Morten Hertzum was supported by a grant from the Danish National Research Foundation. The design and implementation of TeSS as well as the design and execution of the experiment were done in collaboration with Jette Broløs, Marta Lárusdóttir, Kristian Pilgaard, and Flemming Sørensen. We wish to thank the students who participated in the experiment as subjects, and the CHI-reviewers. We are indebted to Per Settergren Sørensen for his support on statistical issues and to Peter Naur for many judicious proposals for clarification.

REFERENCES

1. Bailey, R.W. Performance vs. preference, in *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting* (Seattle WA, October 1993), HFES, 282-285.
2. Bevan, N. Measuring usability as quality of use. *Software Quality Journal* 4 (1995), 115-150.
3. Card, S.K., Moran, T.P. and Newell, A. The keystroke-level model for user performance time with interactive systems. *Communications of the ACM* 23, 7 (1980), 396-410.
4. Cohen, J. and Cohen, P. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale NJ, 1975.
5. Gutwin, C. and Greenberg, S. Effects of awareness support on groupware usability, in *Proceedings of CHI '98* (Los Angeles CA, May 1998), ACM Press, 511-518.
6. Hertzum, M. and Frøkjær, E. Browsing and querying in online documentation: A study of user interfaces and the interaction process. *ACM Transactions on Computer-Human Interaction* 3, 2 (1996), 136-161.
7. ISO 9241-11. Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability (1998).
8. Kirakowski, J. and Corbett, M. SUMI: The software usability measurement inventory. *British Journal of Educational Technology* 24, 3 (1993), 210-212.
9. Naur, P. Intuition in software development, in H. Ehring et al. *Formal Methods and Software Development*, Vol. 2, Lecture Notes in Computer Science 186, Springer, Berlin, 1985. Also in Naur, P. *Computing: A Human Activity*. ACM Press/Addison-Wesley, New York, 1992, 449-466.
10. Newman, W. and Taylor, A. Toward a methodology employing critical parameters to deliver performance improvements in interactive systems, in *Proceedings of INTERACT '99* (Edinburgh, August 1999), IOS Press, 605-612.
11. Nielsen, J. *Usability Engineering*, Academic Press, Boston, 1993.
12. Nielsen, J. and Levy, J. Measuring usability: Preference vs. performance, *Communications of the ACM* 37, 4 (1994), 66-75.
13. Omoigui, N., He, L., Gupta, A., Grudin, J. and Sanocki, E. Time-compression: Systems concerns, usage, and benefits, in *Proceedings of CHI '99* (Pittsburg PA, May 1999), ACM Press, 136-143.
14. Rasmussen, J. Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics* SMC-13, 3 (1983), 257-266.
15. Su, L.T. Evaluation measures for interactive information retrieval. *Information Processing & Management* 28, 4 (1992), 503-516.
16. Van Welie, M., van der Veer, G.C. and Eliëns, A. Breaking Down Usability, in *Proceedings of INTERACT '99* (Edinburgh, August 1999), IOS Press, 613-620.
17. Walker, M.A., Fromer, J., Di Fabbrizio, G., Mestel, C. and Hindle, D. What can I say?: Evaluating a spoken language interface to email, in *Proceedings of CHI '98* (Los Angeles CA, May 1998), ACM Press, 582-589.